University of Stuttgart
Germany

Lexical Semantic Change Discovery

Sinan Kurtyigit
Supervisor: apl. Prof. Dr. Sabine Schulte im Walde
Advisor: Dominik Schlechtweg
Institute for Natural Language Processing, University of Stuttgart, Germany

Introduction

- Words lose or gain sense(s) over time (e.g., Zelle ('cell')).
- The field of Lexical Semantic Change Detection (LSCD) focuses on detecting semantic changes.
- Only a limited amount of work in LSCD focuses on discovering novel instances of semantic change.
- The goal is to make lexical semantic change useful.

Goals

Main goal is to solve the task of **lexical semantic change discovery**:

> Given a corpus pair (C_1, C_2), decide for the intersection of their vocabularies which words lost or gained sense(s) between C_1 and C_2.

Therefore, a framework to automatically discover novel changing words is build.

- Discovery process is fully automated.
- Easily applicable for a wide range of user.
- Additional tools are provided for evaluating and fine-tuning.

Discovery Process

1. Generate word embeddings for words in vocabulary intersection.
2. Measure differences between word embeddings from C_1 and C_2.
3. Calculate a threshold. Mark words with a value greater than or equal to this threshold as changing.
4. Filter out undesirable words.
5. (Optional) Extract usages and store in specific format (DURel).

Models

Two models are provided to generate word embeddings:

1. A static model (SGNS, Mikolov et al., 2013a;b) that generates one word embedding.
2. A contextualized model (BERT, Devlin et al., 2019) that generates one word embedding for every word usage, i.e., a sentence where the word occurs.

Static Approach

Most static approaches in LSCD combine three sub-systems to generate graded values (Schlechtweg et al., 2019):

1. Creating word embeddings.
2. Aligning them across corpora.
3. Measuring differences between the aligned embeddings.

Contextualized Approach

The process of generating graded values is slightly different:

1. Sample words that will act as an input for the model.
2. Extract usages from the corpora for the sampled words.
3. Create two sets of contextualized word embeddings for every word in the sample.
4. Measure differences between the two sets of embeddings.

Thresholding

According to the graded values a threshold is calculated (Kaiser et al., 2020b):

$$TH = \mu + t \cdot \sigma, \tag{1}$$

where μ is the mean and σ standard deviation.
Words whose graded values are greater than or equal to this threshold, are labeled as changing.

Filtering

Two filters are provided to remove undesirable words:

1. A lemma-level filter.
2. (Optional) A usage-level filter.

Store Usages for Human-Annotation

- Usages for the predicted words are extracted and stored in a specific format.
- These can be uploaded to the openly available DURel interface for annotation and visualization.[1]
- Useful for evaluating the quality of the predictions and detecting false positives.

[1] `https://www.ims.uni-stuttgart.de/data/durel-tool`.

Recap

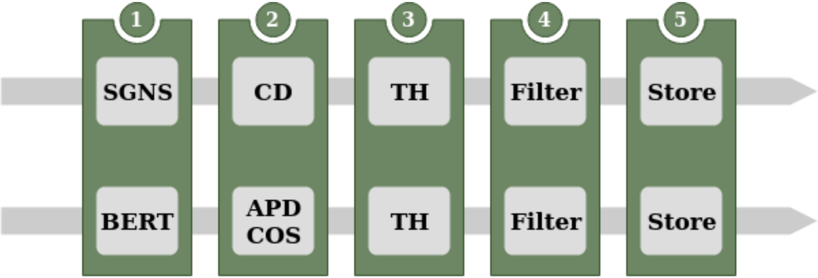

Figure 1: The essential steps of the discovery process.

Framework Applicatioon

- The framework is showcased by an exemplary discovery process on the German SemEval-2020 data set (Schlechtweg et al., 2020).
- Additionally, the model parameters are fine-tuned by solving the SemEval-2020 subtasks.

Data and Subtasks

The data set includes:

- Two time-specific Corpora C_1 (DTA, 1800–1899) und C_2 (BZ+ND 1946–1990).
- 48 target words.
- Binary und graded gold data for evaluation.

Subtasks:

1. Binary Classification: For a set of target words, decide which words lost or gained sense(s) between C_1 and C_2.
2. Graded Ranking: Rank a set of target words according to their degree of LSC between C_1 and C_2.

Tuning

The discovery process is closely related to the SemEval-2020 subtasks.

1. Subtask 2 is solved to optimize the graded value predictions.
2. Afterwards, Subtask 1 is solved to find the best-performing threshold
3. The best parameter configuration for both models are then used to discover changing words.

Annotation

The model predictions are validated by human annotation:[2]

1. Usages are uploaded to the DURel interface for annotation and visualization.
2. Annotators judge the semantic relatedness of pairs of word usages.

4:	Identical
3:	Closely Related
2:	Distantly Related
1:	Unrelated

Table 1: DURel relatedness scale (**?**)

[2]The data set is available at
`https://www.ims.uni-stuttgart.de/data/wugs`

Example

1. Es ist richtig, dass mit dem **Aufkommen** der Manufaktur im Unterschied zum Handwerk sich Spuren der Kinderexploitation zeigen.
   '*It is true that with the **emergence** of the manufactory, in contrast to the handicraft, traces of child exploitation are showing.*'
2. Sie wissen, daß wir für das Vieh mehr Futter aus eigenem **Aufkommen** brauchen.
   '*They know that we need more feed from our own **production** for the cattle.*'

Word Usage Graphs

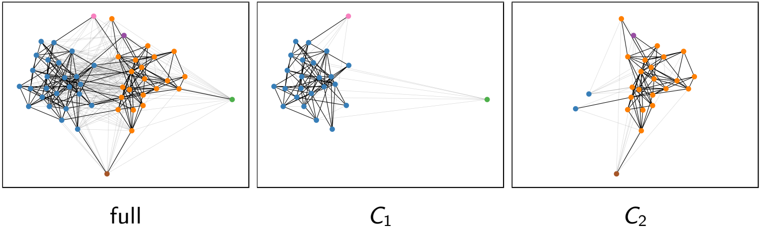

Figure 2: Word Usage Graph of German *Aufkommen* (left), subgraphs for first time period C_1 (middle) and for second time period C_2 (right). **black**/gray lines indicate **high**/low edge weights.

Results

	parameters	t	tuning				predictions			
			ρ	$F_{0.5}$	P	R	ρ	$F_{0.5}$	P	R
SGNS	$k = 1,\ s = .005$	1.0	.690	.692	.750	.529				
	$\mathbf{k = 5, s = .001}$	1.0	.710	**.738**	.818	.529	.324	**.748**	.704	1.0
	$k = 5,\ s =$ None	1.0	.710	.685	.714	.588				
BERT	APD	$-.2$	.673	.598	.560	.824				
	COS	.1	.738	**.741**	.706	.788	.482	.620	.567	1.0

Table 2: Performance (Spearman ρ, $F_{0.5}$-measure, precision P and recall R) of different approaches on tuning data (SemEval targets) as well as performance of best static and contextualized approach on respective predictions with optimal tuning threshold t.

Error Sources

1. **Context Change**: Words where the context in the usages shifts between C_1 and C_2, e.g., *Angriffswaffe* ('offensive weapon'), *aussterben* ('to die out') and *Königreich* ('kingdom').
2. **Context Variety**: Word that can be used in a large variety of contexts, e.g., *neunjährig* ('9-year-old'), *vorjährig* ('of the previous year') and *Bemerken* ('notice').

WUG - Angriffswaffe

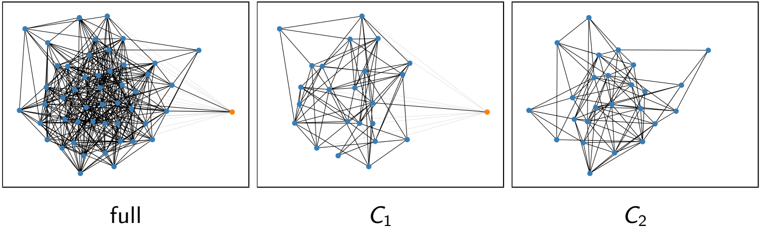

Figure 3: Word Usage Graph of German *Anriffswaffe* (left), subgraphs for first time period C_1 (middle) and for second time period C_2 (right).

Changing Word

1. Man sieht also, daß die Striche nach den Tausenden, nach den Hunderten und nach den **Zehnern** gesetzt werden.
   '*So you can see that the strokes are placed after the thousands, after the hundreds, and after the **tens**.*'
2. Fußball-Toto : Kein Elfer ; 6 **Zehner** mit je 3778 Mark ; 152 Neuner mit je 298 Mark.
   '*Soccer lottery : No eleven ; 6 **tens** with 3778 marks each ; 152 nines with 298 marks each.*'

WUG

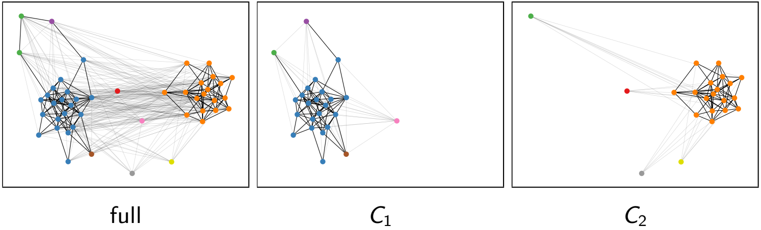

Figure 4: Word Usage Graph of German *Zehner* (left), subgraphs for first time period C_1 (middle) and for second time period C_2 (right).

Conclusion

The goal was to make LSCD useful by providing a framework to automatically discover changing words.

- Both approaches successfully discovered changing words.
- Static model performed better and is recommended.
- The framework was successfully used beyond LSC discovery.

Weaknesses

- Many falsely predicted changing words.
- High-performance on SemEval-2020 data might not translate to other data sets.
- At least small fine-tuning might be necessary.

End

Thank you for your attention.